Selected Ph.D. Thesis Abstracts

This Ph.D. thesis abstracts section presents abstracts of 24 theses defended during the period of January 2016 to July 2017. These abstracts were selected out of 43 submissions covering numerous research and applications under intelligent informatics such as robotics, cyber-physical systems, social intelligence, intelligent agent technology, recommender systems, complex rough sets, multi-view learning, text mining, general-purpose learning systems, bayesian modeling, knowledge graphs, genetic programming and dynamic adversarial mining. These thesis abstracts are organized by the alphabetical order of the author's family name. Fig. 1 presents a word cloud of the submitted abstracts.

Fig. 1. Word Cloud of the Ph.D Abstracts

Inferring Diffusion Models with Structural and Behavioral Dependency in Social Networks

Qing Bao
qingbao@comp.hkbu.edu.hk
Hong Kong Baptist University, Hong Kong

ONLINE social and information networks, like Facebook and Twitter, exploit the in- fluence of neighbors to achieve effective information sharing and spreading. The process that information is spread via the connected nodes in social and information networks is referred to as diffusion. In the literature, a number of diffusion models have been proposed for different applications like influential user identification and personalized recommendation. However, comprehensive studies to discover the hidden diffusion mechanisms governing the information diffusion using the datadriven paradigm are still lacking. This thesis research aims to design novel diffusion models with the structural and behaviorable dependency of neighboring nodes for representing social networks, and to develop computational algorithms to infer the diffusion models as well as the underlying diffusion mechanisms based on information cascades observed in real social networks.

By incorporating structural dependency and diversity of node neighborhood into a widely used diffusion model called Independent Cascade (IC) Model, we first propose a component-based diffusion model where the influence of parent nodes is exerted via connected components. Instead of estimating the node-based diffusion probabilities as in the IC Model, component-based diffusion probabilities are estimated using an expectation maximization (EM) algorithm derived under a Bayesian framework. Also, a newly derived structural diversity measure namely dynamic effective size is proposed for quantifying the dynamic information redundancy within each parent component.

The component-based diffusion model suggests that node connectivity is a good proxy to quantify how a nodes activation behavior is affected by its node neighborhood. To model directly the behavioral dependency of node neighborhood, we then propose a co-activation pattern based diffusion model by integrating the latent class model into the IC Model where the co-activation patterns of parent nodes form the latent classes for each node. Both the co-activation patterns and the corresponding pattern-based diffusion probabilities are inferred using a two-level EM algorithm. As compared to the component-based diffusion model, the inferred co-activation patterns can be interpreted as the soft parent components, providing insights on how each node is influenced by its neighbors as reflected by the observed cascade data.

With the motivation to discover a common set of the over-represented temporal activation patterns (motifs) characterizing the overall diffusion in a social network, we further propose a motif-based diffusion model. By considering the temporal ordering of the parent activations and the social roles estimated for each node, each temporal activation motif is represented using a Markov chain with the social roles being its states. Again, a two-level EM algorithm is proposed to infer both the temporal activation motifs and the corresponding diffusion network simultaneously. The inferred activation motifs can be interpreted as the underlying diffusion mechanisms characterizing the diffusion happening in the social network.

Extensive experiments have been carried out to evaluate the performance of all the proposed diffusion models using both synthetic and real data. The results obtained and presented in the thesis demonstrate the effectiveness of the proposed models. In addition, we discuss in detail how to interpret the inferred co-activation patterns and interaction motifs as the diffusion mechanisms under the context of different real social network data sets (http://repository.hkbu.edu.hk/etd_oa/305/).

Bayesian Modeling for Optimization and Control in Robotics

Roberto Calandra
roberto.calandra@berkeley.edu
UC Berkeley, USA.

ROBOTICS has the potential to be one of the most revolutionary technologies in human history. The impact of cheap and potentially limitless manpower could have a profound influence on our everyday life and overall onto our society. As envisioned by Iain M. Banks, Asimov and many other science fictions writers, the effects of robotics on our society might lead to the disappearance of physical labor and a generalized increase of the quality of life. However, the large-scale deployment of robots in our society is still far from reality, except perhaps in a few niche markets such as manufacturing. One reason for this limited deployment of robots is that, despite the tremendous advances in the capabilities of the robotic hardware, a similar advance on the control software is still lacking. The use of robots in our everyday life is still hindered by the necessary complexity to manually design and tune the controllers used to execute tasks. As a result, the deployment of robots often requires lengthy and extensive validations based on human expert knowledge, which limit their adaptation capabilities and their widespread diffusion. In the future, in order to truly achieve an ubiquitous robotization of our society, it is necessary to reduce the complexity of deploying new robots in new environments and tasks.

The goal of this dissertation is to provide automatic tools based on Machine Learning techniques to simplify and streamline the design of controllers for new tasks. In particular, we here argue that Bayesian modeling is an important tool for automatically learning models from raw data and properly capture the uncertainty of the such models. Automatically learning models however requires the definition of appropriate features used as input for the model. Hence, we present an approach that extend traditional Gaussian process models by jointly learning an appropriate feature representation and the subsequent model. By doing so, we can strongly guide the features representation to be useful for the subsequent prediction task.

A first robotics application where the use of Bayesian modeling is beneficial is the accurate learning of complex dynamics models. For highly non-linear robotic systems, such as in presence of contacts, the use of analytical system identification techniques can be challenging and time-consuming, or even intractable. We introduce a new approach for learning inverse dynamics models exploiting artificial tactile sensors. This approach allows to recognize and compensate for the presence of unknown contacts, without requiring a spatial calibration of the tactile sensors. We demonstrate on the humanoid robot iCub that our approach outperforms state-of-the-art analytical models, and when employed in control tasks significantly improves the tracking accuracy.

A second robotics application of Bayesian modeling is automatic black-box optimization of the parameters of a controller. When the dynamics of a system cannot be modeled (either out of complexity or due to the lack of a full state representation), it is still possible to solve a task by adapting an existing controller. The approach used in this thesis is Bayesian optimization, which allows to automatically optimize the parameters of the controller for a specific task. We evaluate and compare the performance of Bayesian optimization on a gait optimization task on the dynamic bipedal walker Fox. Our experiments highlight the benefit of this approach by reducing the parameters tuning time from weeks to a single day.

In many robotic application, it is however not possible to always define a single straightforward desired objective. More often, multiple conflicting objectives are desirable at the same time, and thus the designer needs to take a decision about the desired trade-off between such objectives (e.g., velocity vs. energy consumption). One framework that is useful to assist in this decision making is the multi-objective optimization framework, and in particular the definition of Pareto optimality. We propose a novel framework that leverages the use of Bayesian modeling to improve the quality of traditional multi-objective optimization approaches, even in low-data regimes. By removing the misleading effects of stochastic noise, the designer is presented with an accurate and continuous Pareto front from which to choose the desired trade-off. Additionally, our framework allows the seamless introduction of multiple robustness metrics which can be considered during the design phase. These contributions allow an unprecedented support to the design process of complex robotic systems in presence of multiple objective, and in particular with regards to robustness.

The overall work in this thesis successfully demonstrates on real robots that the complexity of deploying robots to solve new tasks can be greatly reduced trough automatic learning techniques. We believe this is a first step towards a future where robots can be used outside of closely supervised environments, and where a newly deployed robot could quickly and automatically adapt to accomplish the desired tasks (http://tuprints.ulb.tu-darmstadt.de/5878/13/thesis_roberto_calandra.pdf).

APPLICATION OF AGENT TECHNOLOGY FOR FAULT DIAGNOSIS OF TELECOMMUNICATION NETWORKS

Alvaro Carrera
a.carrera@upm.es
Universidad Politecnica de Madrid, Spain

THIS PhD thesis contributes to the problem of autonomic fault diagnosis of telecommunication networks. Nowadays, in telecommunication networks, operators perform manual diagnosis tasks. Those operations must be carried out by high skilled network engineers which have increasing difficulties to properly manage the growing of those networks, both in size, complexity and heterogeneity. Moreover, the advent of the Future Internet makes the demand of solutions which simplifies and automates the telecommunication network management has been increased in recent years. To collect the domain knowledge required to developed the proposed solutions and to simplify its adoption by the operators, an agile testing methodology is defined for multi-agent systems. This methodology is focused on the communication gap between the different work groups involved in any software development project, stakeholders and developers. To contribute to overcoming the problem of autonomic fault diagnosis, an agent architecture for fault diagnosis of telecommunication networks is defined. That architecture extends the Belief-Desire-Intention (BDI) agent model with different diagnostic

models which handle the different subtasks of the process. The proposed architecture combines different reasoning techniques to achieve its objective using a structural model of the network, which uses ontology-based reasoning, and a causal model, which uses Bayesian reasoning to properly handle the uncertainty of the diagnosis process. To ensure the suitability of the proposed architecture in complex and heterogeneous environments, an argumentation framework is defined. This framework allows agents to perform fault diagnosis in federated domains. To apply this framework in a multi-agent system, a coordination protocol is defined. This protocol is used by agents to dialogue until a reliable conclusion for a specific diagnosis case is reached. Future work comprises the further extension of the agent architecture to approach other managements problems, such as self-discovery or self-optimisation; the application of reputation techniques in the argumentation framework to improve the extensibility of the diagnostic system in federated domains; and the application of the proposed agent architecture in emergent networking architectures, such as SDN, which offers new capabilities of control for the network (http://oa.upm.es/39170/).

MULTI-AGENT SYSTEM FOR COORDINATION OF DISTRIBUTED ENERGY RESOURCES IN VIRTUAL POWER PLANTS

Anders Clausen
ancla@mmmi.sdu.dk
Centre for Smart Energy Solution, University of Southern Denmark, Denmark

THE electricity grid is facing challenges as a result of an increase in the share of renewable energy in electricity production. In this context, Demand Response (DR) is considered an inexpensive and CO_2-friendly approach to handle the resulting fluctuations in the electricity production. The concept of DR refers to changes in consumption patterns of Distributed Energy Resources (DER), in response to incentive payments or changes in the price of electricity over time. Existing DR programs have capacity requirements, which individual DER entities are often unable to meet. As a consequence, the concept of a Virtual Power Plant (VPP) has emerged. A VPP aggregates multiple DER, and exposes them as a single, controllable entity.

The contribution of this thesis is a general method for integrating DER in VPPs. The approach constitutes a meta-model for VPPs, which describes DER and VPPs as entities. Each entity is constituted by a group of software agents. The meta-model describes the interaction between groups and contains two negotiation models: the intradomain- and the interdomain negotiation models. The intradomain negotiation model describes agent decision logic and communication between agents in a group. The model contains a mediator-based negotiation protocol, where agents negotiate over a set of issues, allowing each entity to pursue several objectives and decide upon several issues.

The interdomain negotiation model describes negotiation between groups of agents. In practice this means that the interdomain negotiation model ties instances of intradomain negotiation together. A key aspect of the interdomain negotiation model is that it ensures group autonomy.

Both negotiation models have been implemented in Controleum, a framework for multi-objective optimization. In this context, Controleum has been refactored to allow for the abstractions of the negotiation models to be implemented in the framework. Furthermore, a Domain Specific Language (DSL) has been developed and implemented to allow for easy configuration of negotiations.

Experiments with simulations of different VPP scenarios have been conducted. These experiments indicate that the proposed approach is capable of integrating complex and heterogeneous DER in VPPs, while preserving the autonomous nature of the DER. Experiments are also conducted on instances of the 0/1 Knapsack problem. These experiments serve to illustrate the general applicability of the proposed solution.Future experiments will test the solution in real scenarios (http://aclausen.dk/documents/thesis.pdf).

EFFICIENT KNOWLEDGE MANAGEMENT FOR NAMED ENTITIES FROM TEXT

Sourav Dutta
sourav.dutta@nokia.com
Saarland University, Germany

THE evolution of search from keywords to entities has necessitated the efficient harvesting and management of entity-centric information for constructing knowledge bases catering to various applications such as semantic search, question answering, and information retrieval. The vast amounts of natural language texts available across diverse domains on the Web provide rich sources for discovering facts about named entities such as people, places, and organizations.

A key challenge, in this regard, entails the need for precise identification and disambiguation of entities across documents for extraction of attributes/relations and their proper representation in knowledge bases. Additionally, the applicability of such repositories not only involves the quality and accuracy of the stored information, but also storage management and query processing efficiency. This dissertation aims to tackle the above problems by presenting efficient approaches for entity-centric knowledge acquisition from texts and its representation in knowledge repositories.

This dissertation presents a robust approach for identifying text phrases pertaining to the same named entity across huge corpora, and their disambiguation to canonical entities present in a knowledge base, by using enriched semantic contexts and link validation encapsulated in a hierarchical clustering framework. This work further presents language and consistency features for classification models to compute the credibility of obtained textual facts, ensuring quality of the extracted information. Finally, an encoding algorithm, using frequent term detection and improved data locality, to represent entities for enhanced knowledge base storage and query performance is presented (http://scidok.sulb.uni-saarland.de/volltexte/2017/6792/).

BAYESIAN MODELS OF CATEGORY ACQUISITION AND MEANING DEVELOPMENT

Lea Frermann
l.frermann@sms.ed.ac.uk
University of Edinburgh, UK.

THE ability to organize concepts (e.g., dog, chair) into efficient mental representations, i.e., categories (e.g., animal, furniture) is a fundamental mechanism which allows humans to perceive, organize, and adapt to their world. Much research has been dedicated to the questions of how categories emerge and how they are represented. Experimental evidence suggests that (i) concepts and categories are represented through sets of features (e.g., dogs bark, chairs are made of wood) which are structured into different types (e.g, behavior, material); (ii) categories and their featural representations are learnt jointly and incrementally; and (iii) categories are dynamic and their representations adapt to changing environments.

This thesis investigates the mechanisms underlying the incremental and dynamic formation of categories and their featural representations through cognitively motivated Bayesian computational models. Models of category acquisition have been extensively studied in cognitive science and primarily tested on perceptual abstractions or artificial stimuli. In this thesis, we focus on categories acquired from natural language stimuli, using nouns as a stand-in for their reference concepts, and their linguistic contexts as a representation of the concepts features. The use of text corpora allows us to (i) develop large-scale unsupervised models thus simulating human learning, and (ii) model child category acquisition, leveraging the linguistic input available to children in the form of transcribed child-directed language.

In the first part of this thesis we investigate the incremental process of category acquisition. We present a Bayesian model and an incremental learning algorithm which sequentially integrates newly observed data. We evaluate our model output against gold standard categories (elicited experimentally from human participants), and show that high-quality categories are learnt both from child-directed data and from large, thematically unrestricted text corpora. We find that the model performs well even under constrained memory resources, resembling human cognitive limitations. While lists of representative features for categories emerge from this model, they are neither structured nor jointly optimized with the categories.

We address these shortcomings in the second part of the thesis, and present a Bayesian model which jointly learns categories and structured featural representations. We present both batch and incremental learning algorithms, and demonstrate the models effectiveness on both encyclopedic and child-directed data. We show that high-quality categories and features emerge in the joint learning process, and that the structured features are intuitively interpretable through human plausibility judgment evaluation.

In the third part of the thesis we turn to the dynamic nature of meaning: categories and their featural representations change over time, e.g., children distinguish some types of features (such as size and shade) less clearly than adults, and word meanings adapt to our ever changing environment and its structure. We present a dynamic Bayesian model of meaning change, which infers time-specific concept representations as a set of feature types and their prevalence, and captures their development as a smooth process. We analyze the development of concept representations in their complexity over time from child-directed data, and show that our model captures established patterns of child concept learning. We also apply our model to diachronic change of word meaning, modeling how word senses change internally and in prevalence over centuries.

The contributions of this thesis are threefold. Firstly, we show that a variety of experimental results on the acquisition and representation of categories can be captured with computational models within the framework of Bayesian modeling. Secondly, we show that natural language text is an appropriate source of information for modeling categorization-related phenomena suggesting that the environmental structure that drives category formation is encoded in this data. Thirdly, we show that the experimental findings hold on a larger scale. Our models are trained and tested on a larger set of concepts and categories than is common in behavioral experiments and the categories and featural representations they can learn from linguistic text are in principle unrestricted (http://frermann.de/dataFiles/phd_thesis_leafrermann.pdf).

LATENT FACTOR MODELS FOR COLLABORATIVE FILTERING

Anupriya Gogna
anupriyag@iiitd.ac.in
Indraprastha Institute of Information Technology - Delhi, India

THE enormous growth in online availability of information content has made Recommender Systems (RS) an integral part of most online portals and e-commerce sites. Most websites and service portals, be it movie rental services, online shopping or travel package providers, offer some form of recommendations to users. These recommendations provide the users more clarity, that too expeditiously and accurately in limiting (shortlisting) the items/information they need to search through, thereby improving the customer's experience. The direct link between customer's satisfaction and revenue of e-commerce sites induce widespread interest of both, academia and industry, in the design of efficient recommender systems.

The current de-facto approach for RS design is Collaborative Filtering (CF). CF techniques use the ratings provided by users, to a subset of the items in the repository, to make future recommendations. However, the rating information is hard to acquire; often a user has rated less than 5% of the items. Thus, the biggest challenge in recommender system design is to infer users preference from this extremely limited predilection information. The lack of adequate (explicit) preference information has motivated several works to augment the rating data with auxiliary information such as users demographics, trust networks, and item tags. Further, the scale of the problem, i.e. the amount of the data to be processed (selecting few items out of hundreds and thousands of items for an equally large number of users) adds another dimension to the concerns

surrounding the design of a good RS. There have been several developments in the field of RS design over the past decades. However, the difficulty in achieving the desired accuracy and effectiveness in recommendations leaves considerable scope for improvement.

In this work, we model effective recommendation strategies, using optimization centric frameworks, by exploiting reliable and readily available information, to address several pertinent issues concerning RS design. Our proposed recommendation strategies are built on the principals of latent factor models (LFM). LFM are constructed on the belief that a users choice for an item is governed by a handful of factors C the latent factors. For example, in the case of movies, these factors may be genre, director, language while for hotels it can be price and location.

Our first contribution targets improvement in prediction accuracy as well the speed of processing by suggesting modifications to the standard LFM frameworks. We develop a more intuitive model, supported by effective algorithm design, which better captures the underlying structure of the rating database while ensuring a reduction in run time compared to standard CF techniques. In the next step, we build upon these proposed frameworks to address the problem of lack of collaborative data, especially for cold start (new) users and items, by making use of readily available user and item metadata - item category and user demographics. Our suggested frameworks make use of available metadata to add additional constraints in the standard models; thereby presenting a comprehensive strategy to improve prediction accuracy in both warm (existing users/items for which rating data is available) and cold start scenario.

Although, high recommendation accuracy is the hallmark of a good RS, over-emphasis on accuracy compromises on variety and leads to monotony. Our next set of models aims to address this concern and promote diversity and novelty in recommendations. Most existing works, targeting diversity, build ad-hoc exploratory models relying heavily on heuristic formulations. In the proposed work, we modify the latent factor model to formulate a joint optimization strategy to establish accuracy-diversity balance; our models yield superior results than existing works.

The last contribution of this work is to explore the use of another representation learning tool for collaborative filtering C Autoencoder (AE). Conventional AE based designs, use only the rating information; lack of adequate data hampers the performance of these structures, thus, they do not perform as well as conventional LFM based designs. In this work, we propose a modification of the standard autoencoder C the Supervised Autoencoder C which can jointly accommodate information from multiple sources resulting in better performance than existing architectures (https://repository.iiitd.edu.in/xmlui/handle/123456789/501).

MACHINE LEARNING THROUGH EXPLORATION FOR PERCEPTION-DRIVEN ROBOTICS

Herke van Hoof
herke.vanhoof@mail.mcgill.ca
McGill University, Canada

THE ability of robots to perform tasks in human environments, such as our homes, has largely been limited to rather simple tasks, such as lawn mowing and vacuum cleaning. One reason for this limitation is that every home has a different layout with different objects and furniture. Thus, it is impossible for a human designer to anticipate all challenges a robot might face, and equip the robot a priori with all necessary perceptual and manipulation skills.

Instead, robots could use machine learning techniques to adapt to new environments. Many current learning techniques, however, rely on human supervisors to provide data in the form of annotations, demonstrations, and parameter settings. As such, making a robot perform a task in a novel environment can still require a significant time investment. In this thesis, instead, multiple techniques are studied to let robots collect their own training data through autonomous exploration.

The first study concerns an unsupervised robot that learns from sensory feedback obtained though interactive exploration of its environment. In a novel bottom-up, probabilistic approach, the robot tries to segment the objects in its environment through clustering with minimal prior knowledge. This clustering is based on cues elicited through the robots actions. Evaluations on a real robot system show that the proposed method handles noisy inputs better than previous methods. Furthermore, a proposed scheme for action selection criterion according to the expected information gain criterion is shown to increase the learning speed.

Often, however, the goal of a robot is not just to learn the structure of the environment, but to learn how to perform a task encoded by a reward signal. In a second study, a novel robot reinforcement learning algorithm is proposed that uses learned non-parametric models, value functions, and policies that can deal with high-dimensional sensory representations. To avoid that the robot converges prematurely to a sub-optimal solution, the information loss of policy updates is limited. The experiments show that the proposed algorithm performs well relative to prior methods. Furthermore, the method is validated on a real-robot setup with high-dimensional camera image inputs.

One problem with typical exploration strategies is that the behavior is perturbed independently in each time step. The resulting incoherent exploration behavior can result in inefficient random walk behavior and wear and tear on the robot. Perturbing policy parameters for an entire episode yields coherent exploration, but tends to increase the number of episodes needed. In a third study, a strategy is introduced that makes a balanced trade-off between these two approaches. The experiments show that such trade-offs are beneficial across different tasks and learning algorithms.

This thesis thus addresses how robots can learn autonomously by exploring the world. Throughout the thesis, new approaches and algorithms are introduced: a probabilistic interactive segmentation approach, the non-parametric relative entropy policy search algorithm, and a framework for generalized exploration. These approaches and algorithms contribute towards the capability of robots to autonomously

learn useful skills in human environments in a practical manner (http://tuprints.ulb.tu-darmstadt.de/5749/).

A Service-Oriented Framework for the Specification, Deployment, Execution, Benchmarking, and Prediction of Performance of Scalable Privacy-Preserving Record Linkage Techniques

Dimitrios Karapiperis
dkarapiperis@eap.gr
Hellenic Open University, Greece

AT the dawn of a new era of computing and the growth of big data, information integration is more important than ever before. Large organizations, such as corporations, health providers, public sector agencies, or research institutes, integrate their data in order to generate insightful data analytics. This data integration and analysis enables these organizations to make certain decisions toward deriving better business outcomes.

Record linkage, also known as entity resolution or data matching, is the process of resolving whether two records that belong to disparate data sets, refer to the same real- world entity. The lack of common identifiers, as well as typos and inconsistencies in the data, render the process of record linkage very challenging and mandatory for organizations which need to integrate their records. When data is deemed as private, then specialized techniques are employed that perform Privacy-Preserving Record Linkage (PPRL) in a secure manner, by respecting the privacy of the individuals who are represented by those records. For instance, in the public sector, there are databases which contain records that refer to the same citizen holding outdated information. Although, there is an urgent need of integration, the lack of common identifiers poses significant impediments in the linkage process.

Due to the large volumes of records contained in the data sets, core component of PPRL is the blocking phase, in which records are inserted into overlapping blocks and, then, are compared with one another. The purpose of the blocking phase is to formulate as many as possible matching record pairs. The blocking methods proposed thus far in the literature apply empirical techniques, which, given the particularities and technical characteristics of the data sets at hand, produce arbitrary results. This dissertation is the first to provide theoretical guarantees of completeness in the generated result set of the PPRL process, by introducing a randomized framework that allows for easy tuning of its configuration. Its flexibility lies in the fact that one can specify the level of its performance, with respect to the completeness of the results, taking into account multiple factors, such as: the urgency of the problem being solved, the desired response time, or the criticalness of the results completeness. Additionally, we enhance its main functionality, by providing certain extensions, and illustrate its applicability to both offline and online settings. The framework has been materialized by a prototype that is freely available so that it can be used by practitioners and researchers in their tasks.

This dissertation is divided into several chapters; we first introduce the core of our framework and its capabilities, and, then, we present its several extensions, such as the integration with the map/reduce paradigm for scaling up large volumes of records, or the add-on for performing PPRL using numeric values. In each of these chapters, we report on an extensive evaluation of the application of the constituent methods with real data sets, which illustrates that they outperform existing approaches (https://drive.google.com/file/d/0B2tBkOmLy8WZc0Z4OU0zQ2dOMVE/view?usp=sharing).

Learning with Multiple Views of Data: Some Theoretical and Practical Contributions

Chao Lan
pete.chaolan@gmail.com
University of Kansas, USA.

IN machine learning applications, instances are often describable by multiple feature sets, each somewhat interpretable and sufficient for the learning task. In this case, each feature set is called a view, the instances are called multi-view data, and the study of learning with multi-view data is called multi-view learning. In this dissertation, we investigated several issues of multi-view learning in two settings: one is called statistical setting, where one aims to learn predictive models based on random multi-view sample; the other one is called matrix setting, where one aims to recover missing values in the feature matrices of multiple views. In statistical setting, we first theoretically investigated the possibility of training an accurate predictive model using as few unlabeled multi-view data as possible, and concluded such possibility by improving the state-of-the-art unlabeled sample complexity of semi-supervised multi-view learning by a logarithm factor. We then investigated the application of multi-view clustering methods in social circle detection on ego social networks. We finally proposed a simple multi-view multi-class learning algorithm that consistently outperforms the state-of-the-art algorithm. In matrix setting, we focused on the negative transfer problem in Collective Matrix Factorization (CMF), which is a popular method to recover missing values in multi-view feature matrices. We first theoretically characterized negative transfer in a CMF estimator, as the decrease of its ideal minimax learning rate by a root function. We then showed our presented ideal rate is tight (up to a constant factor), by employing the statistical PAC theory to derive a matching upper bound for it; our employment of the PAC theory improved the state-of-the-art one, by relaxing its strong i.i.d. assumption of matrix recovery errors. We finally proposed a simple variant of CMF that outperforms a current variant in small sample case. At the conceptual level, we have been bridging gaps between research in statistical setting and matrix setting. Specifically, we employed statistical PAC theory to analyze matrix recovery error in both active and passive learning settings; we employed statistical multi-view learning framework to develop a variant of CMF for matrix recovery; we employed statistical mini-max theory to analyze CMF performance.

Finally, our research in multi-view learning has motivated two other studies. One study challenged a common assumption in cheminformatics that unreported substance-compound binding profiles are all negative. The other study proposed a first active matrix recovery method with PAC guarantee.
(https://www.dropbox.com/s/1zf3qksqoc8oqnr/dissertation.pdf ?dl=0)

TRANSFORMATION-BASED COMMUNITY DETECTIONFROM SOCIAL NETWORKS

Sungsu Lim
ssungssu@kaist.ac.kr
Korea Advanced Institute of Science and Technology, Korea

IN recent decades social network analysis has become one of the most attractive issues in data mining. In particular, community detection is a fundamental problem in social network analysis. Many theories, models, and methods have been developed for this purpose. However, owing to a wide variety of network structures, there remain challenges to determining community structures from social networks. Among them, we focus on solving four important problems for community detection with different underlying structures. These are identifying the community structure of a graph when it consists of (i) overlapping community structure, (ii) highly mixed community structure, (iii) complex sub-structure, and (iv) highly mixed overlapping community structure.

We address these problems by developing transformation techniques. Our key motivation is that the transformation of a given network can provide us an improved structure to identify the community structure of an original network, as a kernel trick does. We propose a transformation-based algorithm that converts an original graph to a transformed graph that reflects the structure of the original network and has a superior structure for each problem. We identify the community structure using the transformed graph and then the membership on the transformed graph is translated back to that of the original graph.

For the first problem, we present a notion of the linkspace transformation that enables us to combine the advantages of both the original graph and the line graph, thereby conveniently achieving overlapping communities. Based on this notion, we develop an overlapping community detection algorithm LinkSCAN* [1]. The experimental results demonstrate that the proposed algorithm outperforms existing algorithms, especially for networks with many highly overlapping nodes.

For the second problem, we propose a community detection algorithm BlackHole [2]. The proposed graph embedding model attempts to place the vertices of the same community at a single position, similar to how a black hole attracts everything nearby. Then, these positions are easily grouped by a conventional clustering algorithm. The proposed algorithm is proven to achieve extremely high performance regardless of the difficulty of community detection in terms of the mixing and the clustering coefficient.

For the third problem, we propose a motif-based embedding method for graph clustering by modeling higher-order relationships among vertices in a graph [3]. The proposed method considers motif-based attractive forces to enhance the clustering tendency of points in the output space of graph embedding. We prove the relationship with graph clustering and verify the performance and applicability.

For the fourth problem, we develop an algorithm to address the highly mixed overlapping community detection problem. The transformation of the proposed algorithm LinkBlackHole consists of a sequence of two different transformations. By first applying the link-space transformation and then the BlackHole transformation, we can detect highly mixed link communities.

The strength of this dissertation is based on a wide variety of community detection problems from social networks. We believe that this work will enhance the quality of community discovery for social network analysis (http: //library.kaist.ac.kr/thesis02/2016/2016D020115245_S1.pdf).

INCREMENTAL AND DEVELOPMENTAL PERSPECTIVES FOR GENERAL-PURPOSE LEARNING SYSTEMS

Fernando Martłnez-Plumed
fmartinez@dsic.upv.es
Universitat Politcnica de Valencia, Spain

THE stupefying success of Artificial Intelligence (AI) for specific problems, from recommender systems to self-driving cars, has not yet been matched with a similar progress in general AI systems, coping with a variety of (different) problems. This dissertation deals with the long-standing problem of creating more general AI systems, through the analysis of their development and the evaluation of their cognitive abilities.

Firstly, this thesis contributes with a general-purpose declarative learning system gErl [2],[3] that meets several desirable characteristics in terms of expressiveness, comprehensibility and versatility. The system works with approaches that are inherently general: inductive programming and reinforcement learning. The system does not rely on a fixed library of learning operators, but can be endowed with new ones, so being able to operate in a wide variety of contexts. This flexibility, jointly with its declarative character, makes it possible to use the system as an instrument for better understanding the role (and difficulty) of the constructs that each task requires.

Secondly, the learning process is also overhauled with a new developmental and lifelong approach for knowledge acquisition, consolidation and forgetting, which is necessary when bounded resources (memory and time) are considered. In this sense we present a parametrisable (hierarchical) approach [4] for structuring knowledge (based on coverage) which is able to check whether the new learnt knowledge can be considered redundant, irrelevant or inconsistent with the old one, and whether it may be built upon previously acquired knowledge.

Thirdly, this thesis analyses whether the use of more ability-oriented evaluation techniques for AI (such as intelligence tests) is a much better alternative to most task-oriented evaluation approaches in AI. Accordingly, we make a review of what has been done when AI systems have been confronted against tasks taken from intelligence tests [5]. In this regard, we scrutinise what intelligence tests measure in machines,

whether they are useful to evaluate AI systems, whether they are really challenging problems, and whether they are useful to understand (human) intelligence. Our aim here is to contribute to a more widespread realisation that more general classes of problems are needed when constructing benchmarks for AI evaluation.

As a final contribution, we show that intelligence tests can also be useful to examine concept dependencies (mental operational constructs) in the cognitive development of artificial systems, therefore supporting the assumption that, even for fluid intelligence tests, the difficult items require a more advanced cognitive development than the simpler ones. In this sense, in [3] we show how several fluid intelligence test problems are addressed by our general-purpose learning system gErl, which, although it is not particularly designed on purpose to solve intelligence tests, is able to perform relatively well for this kind of tests. gErl makes it explicitly how complex each pattern is and what operators are used for each problem, thus providing useful insight into the characteristics and usefulness of these tests when assessing the abilities and cognitive development of AI systems.

Summing up, this dissertation represents one step forward in the hard and long pursuit of making more general AI systems and fostering less customary (and challenging) ability-oriented evaluation approach.

Novel Methods of Measuring the Similarity and Distance between Complex Fuzzy Sets

Josie McCulloch
josie.mcculloch@nottingham.ac.uk
University of Nottingham, UK.

THIS thesis develops measures that enable comparisons of subjective information that is represented through fuzzy sets. Many applications rely on information that is subjective and imprecise due to varying contexts and so fuzzy sets were developed as a method of modelling uncertain data. However, making relative comparisons between data-driven fuzzy sets can be challenging. For example, when data sets are ambiguous or contradictory, then the fuzzy set models often become non-normal or non-convex, making them difficult to compare.

This thesis presents methods of comparing data that may be represented by such (complex) non-normal or non-convex fuzzy sets. The developed approaches for calculating relative comparisons also enable fusing methods of measuring similarity and distance between fuzzy sets. By using multiple methods, more meaningful comparisons of fuzzy sets are possible. Whereas if only a single type of measure is used, ambiguous results are more likely to occur.

This thesis provides a series of advances around the measuring of similarity and distance. Based on them, novel applications are possible, such as personalised and crowd-driven product recommendations. To demonstrate the value of the proposed methods, a recommendation system is developed that enables a person to describe their desired product in relation to one or more other known products. Relative comparisons are then used to find and recommend something that matches a persons subjective preferences. Demonstrations illustrate that the proposed method is useful for comparing complex, nonnormal and non-convex fuzzy sets. In addition, the recommendation system is effective at using this approach to find products that match a given query (http://eprints.nottingham.ac.uk/id/eprint/33401).

Multi-Document Summarization based on Patterns with Wildcards and Probabilistic Topic Modeling

Jipeng Qiang
qjp2100@163.com
Yangzhou University, China

WITH the rapid development of information technology, a huge amount of electronic documents are available online, such as Web news, scientific literature, digital books, email, microblogging, and etc. How to effectively organize and manage such vast amount of text data, and make the system facilitate and show the information to users, have become challenges in the field of intelligent information processing. Therefore, now more than ever, users need access to robust text summarization systems, which can effectively condense information found in a large amount of documents into a short, readable synopsis, or summary. In recent years, with the rapid development of e-commerce and social networks, we can obtain a large amout of short texts, e.g., book reviews, movie reviews, online chatting, and product introductions. A short text probably contains a lot of useful information that can help to learn hidden topics among texts. Meanwhile, only very limited word co-occurrence information is available in short texts compared with long texts, so traditional multi-document summarization algorithms cannot work very well on these texts. Thus, how to generate a summary from multiple documents has important research and practical values.

In this thesis, we study multi-document summarization (MDS) on long texts and multi-document summarization on short texts, and propose several multi-document summarization algorithms based on patterns with wildcards and probability topic modeling. Our main contributions are as follows.

(1) A novel pattern-based model for generic multi-document summarization is proposed. There are two main categories of multi-document summarization: term-based and ontology-based meth- ods. A term-based method cannot deal with the problems of polysemy and synonymy. An ontology-based approach addresses such problems by taking into account of the semantic information of document con- tent, but the construction of ontology requires lots of manpower. To overcome these open problems, this paper presents a pattern-based model for generic multi-document summarization, which exploits closed patterns to extract the most salient sentences from a document collection and reduce redundancy in the summary. Our method calculates the weight of each sentence of a document collection by accumulating the weights of its covering closed patterns with respect to this sentence, and iteratively selects one sentence that owns the highest weight and less similarity to the previously selected sentences, un- til reaching the length limitation. Our method combines

the advantages of the term-based and ontology-based models while avoiding their weaknesses. Empirical studies on the benchmark DUC2004 datasets demonstrate that our pattern-based method significantly outperforms the state-of-the-art methods.

(2) A new MDS paradigm called user-aware multi-document summarization is proposed. The aim of MDS meets the demands of users, and the comments contain implicit information of their care. Therefore, the generated summaries from the reports for an event should be salient according to not only the reports but also the comments. Recently, Bayesian models have successfully been applied to multi-document summarization showing state-of-the-art results in summarization competitions. News articles are often long. Tweets and news comments can be short texts. In this thesis, the corpus includes both short texts and long texts, referred to as heterogeneous text. Long text topic modeling views texts as a mixture of probabilistic topics, and short text topic modeling adopts simple assumption that each text is sampled from only one latent topic. For heterogeneous texts, in this case neither method developed for only long texts nor methods for only short texts can generate satisfying results. In this thesis, we present an innovative method to discover latent topics from a heterogeneous corpus including both long and short texts. Then, we apply the learned topics to the generation of summarizations. Experiments on real-world datasets validate the effectiveness of the proposed model in comparison with other state-of-the-art models.

(3) A new short text topic model based on word embeddings is proposed. Existing methods such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) cannot solve this problem very well since only very limited word co-occurrence information is available in short texts. Based on recent results in word embeddings that learn semantical representations for words from a large corpus, we introduce a novel method, Embedding-based Topic Modeling (ETM), to learn latent topics from short texts. ETM not only solves the problem of very limited word co-occurrence information by aggregating short texts into long pseudo-texts, but also utilizes a Markov Random Field regularized model that gives correlated words a better chance to be put into the same topic. Experiments on real-world datasets validate the effectiveness of our model comparing with the state-of-the-art models (https://drive.google.com/file/d/0B3BkztuizBiyeTltMmFlci10dUU/view?usp=sharing).

PROBABILISTIC MODELS FOR LEARNING FROM CROWDSOURCED DATA

Filipe Rodrigues
rodr@dtu.dk
Denmark Technical University, Denmark

THIS thesis leverages the general framework of probabilistic graphical models to develop probabilistic approaches for learning from crowdsourced data. This type of data is rapidly changing the way we approach many machine learning problems in different areas such as natural language processing, computer vision and music. By exploiting the wisdom of crowds, machine learning researchers and practitioners are able to develop approaches to perform complex tasks in a much more scalable manner. For instance, crowdsourcing platforms like Amazon mechanical turk provide users with an inexpensive and accessible resource for labeling large datasets efficiently. However, the different biases and levels of expertise that are commonly found among different annotators in these platforms deem the development of targeted approaches necessary.

With the issue of annotator heterogeneity in mind, we start by introducing a class of latent expertise models which are able to discern reliable annotators from random ones without access to the ground truth, while jointly learning a logistic regression classifier or a conditional random field. Then, a generalization of Gaussian process classifiers to multiple-annotator settings is developed, which makes it possible to learn non-linear decision boundaries between classes and to develop an active learning methodology that is able to increase the efficiency of crowdsourcing while reducing its cost. Lastly, since the majority of the tasks for which crowdsourced data is commonly used involves complex high-dimensional data such as images or text, two supervised topic models are also proposed, one for classification and another for regression problems. Using real crowdsourced data from Mechanical Turk, we empirically demonstrate the superiority of the aforementioned models over state-ofthe-art approaches in many different tasks such as classifying posts, news stories, images and music, or even predicting the sentiment of a text, the number of stars of a review or the rating of movie.

But the concept of crowdsourcing is not limited to dedicated platforms such as Mechanical Turk. For example, if we consider the social aspects of the modern Web, we begin to perceive the true ubiquitous nature of crowdsourcing. This opened up an exciting new world of possibilities in artificial intelligence. For instance, from the perspective of intelligent transportation systems, the information shared online by crowds provides the context that allows us to better understand how people move in urban environments. In the second part of this thesis, we explore the use of data generated by crowds as additional inputs in order to improve machine learning models. Namely, the problem of understanding public transport demand in the presence of special events such as concerts, sports games or festivals, is considered. First, a probabilistic model is developed for explaining non-habitual overcrowding using crowd-generated information mined from the Web. Then, a Bayesian additive model with Gaussian process components is proposed. Using real data from Singapores transport system and crowd-generated data regarding special events, this model is empirically shown to be able to outperform state-of-the-art approaches for predicting public transport demand. Furthermore, due to its additive formulation, the proposed model is able to breakdown an observed time-series of transport demand into a routine component corresponding to commuting and the contributions of individual special events.

Overall, the models proposed in this thesis for learning from crowdsourced data are of wide applicability and can

be of great value to a broad range of research communities (http://www.fprodrigues.com/thesis_phd_fmpr.pdf).

Exploring Mixed Reality in Distributed Collaborative Learning Environments

Anasol Pena Rios
acpena@essex.ac.uk
University of Essex, UK.

SOCIETY is moving rapidly towards a world, where technology enables people to exist in a blend of physical and virtual realities. In education, this vision involves technologies ranging from smart classrooms to e-learning, creating greater opportunities for distance learners, bringing the potential to change the fundamental nature of universities. However, to date, most online educational platforms have focused on conveying information rather than supporting collaborative physical activities which are common in university science and engineering laboratories. Moreover, even when online laboratory support is considered, such systems tend to be confined to the use of simulations or pre-recorded videos. The lack of support for online collaborative physical laboratory activities, is a serious shortcoming for distance learners and a significant challenge to educators and researchers.

In working towards a solution to this challenge, this thesis presents an innovative mixed reality framework (computational model, conceptual architecture and proof-of-concept implementation) that enables geographically dispersed learners to perform co-creative teamwork using a computer-based prototype comprising hardware and software components.

Contributions from this work include a novel distributed computational model for synchronising physical objects and their 3D virtual representations, expanding the dual-reality paradigm from single linked pairs to complex groupings, addressing the challenge of interconnecting geographically dispersed environments; and the creation of a computational paradigm that blends a model of distributed learning objects with a constructionist pedagogical model, to produce a solution for distributed mixed reality laboratories.

By way of evidence to support the research findings, this thesis reports on evaluations performed with students from eight different universities in six countries, namely China, Malaysia, Mexico, UAE, USA and UK; providing an important insight to the role of social interactions in distance learning, and demonstrating that the inclusion of a physical component made a positive difference to students learning experience, supporting the use of mixed reality objects in educational activities (http://repository.essex.ac.uk/16172/).

Dynamic Adversarial Mining - Effectively applying machine learning in adversarial non-stationary environments

Tegjyot Singh Sethi
t0seth01@louisville.edu
University of Louisville, USA.

WHILE understanding of machine learning and data mining is still in its budding stages, the engineering applications of the same has found immense acceptance and success. Cybersecurity applications such as intrusion detection systems, spam filtering, and CAPTCHA authentication, have all begun adopting machine learning as a viable technique to deal with large scale adversarial activity. However, the naive usage of machine learning in an adversarial setting is prone to reverse engineering and evasion attacks, as most of these techniques were designed primarily for a static setting. The security domain is a dynamic landscape, with an ongoing never ending arms race between the system designer and the attackers. Any solution designed for such a domain needs to take into account an active adversary and needs to evolve over time, in the face of emerging threats. We term this as the Dynamic Adversarial Mining problem, and the presented work provides the foundation for this new interdisciplinary area of research, at the crossroads of Machine Learning, Cybersecurity, and Streaming Data Mining.

We start with a white hat analysis of the vulnerabilities of classification systems to exploratory attack. The proposed Seed-Explore-Exploit framework provides characterization and modeling of attacks, ranging from simple random evasion attacks to sophisticated reverse engineering. It is observed that, even systems having prediction accuracy close to 100%, can be easily evaded with more than 90% precision. This evasion can be performed without any information about the underlying classifier, training dataset, or the domain of application.

Attacks on machine learning systems cause the data to exhibit non stationarity (i.e., the training and the testing data have different distributions). It is necessary to detect these changes in distribution, called concept drift, as they could cause the prediction performance of the model to degrade over time. However, the detection cannot overly rely on labeled data to compute performance explicitly and monitor a drop, as labeling is expensive and time consuming, and at times may not be a possibility altogether. As such, we propose the Margin Density Drift Detection (MD3) algorithm, which can reliably detect concept drift from unlabeled data only. MD3 provides high detection accuracy with a low false alarm rate, making it suitable for cybersecurity applications; where excessive false alarms are expensive and can lead to loss of trust in the warning system. Additionally, MD3 is designed as a classifier independent and streaming algorithm for usage in a variety of continuous never-ending learning systems.

We then propose a Dynamic Adversarial Mining based learning framework, for learning in non stationary and adversarial environments, which provides security by design. The proposed Predict-Detect classifier framework, aims to provide: robustness against attacks, ease of attack detection using unlabeled data, and swift recovery from attacks. Ideas of feature hiding and obfuscation of feature importance are proposed as strategies to enhance the learning frameworks security. Metrics for evaluating the dynamic security of a system and recover-ability after an attack are introduced to provide a practical way of measuring efficacy of dynamic security strategies. The framework is developed as a streaming

data methodology, capable of continually functioning with limited supervision and effectively responding to adversarial dynamics.

The developed ideas, methodology, algorithms, and experimental analysis, aim to provide a foundation for future work in the area of Dynamic Adversarial Mining, wherein a holistic approach to machine learning based security is motivated.

Leveraging Multimodal Information in Semantics and Sentics Analysis of User-Generated Content

Rajiv Shah
rajivshah@smu.edu.sg
Singapore Management University, Singapore

THE amount of user-generated multimedia content (UGC) has increased rapidly in recent years due to the ubiquitous availability of smartphones, digital cameras, and affordable network infrastructures. To benefit people from an automatic semantics and sentics understanding of UGC, this thesis1,2 focuses on developing effective algorithms for several significant multimedia analytics problems. Sentics are common affective patterns associated with natural language concepts exploited for tasks such as emotion recognition from text/speech or sentiment analysis. Knowledge structures derived from UGC are beneficial in an efficient multimedia search, retrieval, and recommendation. However, real-world UGC is complex, and extracting the semantics and sentics from only multimedia content is very difficult because suitable concepts may be exhibited in different representations. Advancements in technology enable users to collect a significant amount of contextual information (e.g., spatial, temporal, and preferential information). Thus, it necessitates analyzing UGC from multiple modalities to facilitate problems such as multimedia summarization, tag ranking and recommendation, preference-aware multimedia recommendation, and multimedia-based e-learning.

We focus on the semantics and sentics understanding of UGC leveraging both content and contextual information. First, for a better semantics understanding of an event from a large collection of photos, we present the EventBuilder system [2]. It enables people to automatically generate a summary for the event in real-time by visualizing different social media such as Wikipedia and Flickr. In particular, we exploit Wikipedia as the event background knowledge to obtain more contextual information about the event. This information is very useful in effective event detection. Next, we solve an optimization problem to produce text summaries for the event. Subsequently, we present the EventSensor system [6] that aims to address sentics understanding and produces a multimedia summary for a given mood. It extracts concepts and mood tags from the visual content and textual metadata of photos and exploits them in a sentics-based multimedia summary. Moreover, we focus on computing tag relevance for UGIs. Specifically, we leverage personal and social contexts of UGIs and follow a neighbor voting scheme to predict and rank tags [1, 5]. Furthermore, we focus on semantics and sentics understanding from videos (http://dl.acm.org/citation.cfm?id=2912032).

Big Data for Social Sciences: Measuring patterns of human behavior through large-scale mobile phone data

Pal Sundsoy
sundsoy@gmail.com
University of OSLO, Norway

ANALYSIS of large amounts of data, so called Big Data, is changing the way we think about science and society. One of the most promising rich Big Data sources is mobile phone data, which has the potential to deliver near real-time information of human behaviour on an individual and societal scale. Several challenges in society can be tackled in a more efficient way if such information is applied in a useful manner. Through seven publications this dissertation shows how anonymized mobile phone data can contribute to the social good and provide insights into human behaviour on a largescale.

The size of the datasets analysed ranges from 500 million to 300 billion phone records, covering millions of people. The key contributions are two-fold: Big Data for Social Good: Through prediction algorithms the results show how mobile phone data can be useful to predict important socio-economic indicators, such as income, illiteracy and poverty in developing countries. Such knowledge can be used to identify where vulnerable groups in society are, improve allocation of resources for poverty alleviation programs, reduce economic shocks, and is a critical component for monitoring poverty rates over time. Further, the dissertation demonstrates how mobile phone data can be used to better understand human behaviour during large shocks and disasters in society, exemplified by an analysis of data from the terror attack 22nd July 2011 in Norway and a natural disaster on the south-coast in Bangladesh. This work leads to an increased understanding of how information spreads, and how millions of people move around. The intention is to identify displaced people faster, cheaper and more accurately than existing survey-based methods.

Big Data for efficient marketing: Finally, the dissertation offers an insight into how anonymised mobile phone data can be used to map out large social networks, covering millions of people, to understand how products spread inside these networks. Results show that by including social patterns and machine learning techniques in a large-scale marketing experiment in Asia, the adoption rate is increased by 13 times compared to the approach used by experienced marketers. A data-driven and scientific approach to marketing, through more tailored campaigns, contributes to less irrelevant offers for the customers, and better cost efficiency for the companies (https://www.duo.uio.no/handle/10852/55139).

TOWARD INTELLIGENT CYBER-PHYSICAL SYSTEMS: ALGORITHMS, ARCHITECTURES, AND APPLICATIONS

Bo Tang
btang@ele.uri.edu
University of Rhode Island, USA.

CYBER-physical systems (CPS) are the new generation of engineered systems integrated with computation and physical processes. The integration of computation, communication and control adds new capabilities to the systems being able to interact with physical world. The uncertainty in physical environment makes future CPS to be more reliant on machine learning algorithms which can learn and accumulate knowledge from historical data to support intelligent decision making. Such CPS with the incorporation of intelligence or smartness are termed as intelligent CPS which are safer, more reliable and more efficient.

This thesis studies fundamental machine learning algorithms in supervised and unsupervised manners and examines new computing architecture for the development of next generation CPS. Two important applications of CPS, including smart pipeline and smart grid, are also studied in this thesis. Particularly, regarding supervised machine learning algorithms, several generative learning and discriminative learning methods are proposed to improve learning performance. For the generative learning, we build novel classification methods based on exponentially embedded families (EEF), a new probability density function (PDF) estimation method, when some of the sufficient statistics are known. For the discriminative learning, we develop an extended nearest neighbor (ENN) method to predict patterns according to the maximum gain of intra-class coherence. The new method makes a prediction in a "two-way communication" style: it considers not only who are the nearest neighbors of the test sample, but also who consider the test sample as their nearest neighbors. By exploiting the generalized class-wise statistics from all training data, the proposed ENN is able to learn from the global distribution, therefore improving pattern recognition performance and providing a powerful technique for a wide range of data analysis applications. Based on the concept of ENN, an anomaly detection method is also developed in an unsupervised manner.

CPS usually have high-dimensional data, such as text, video, and other multi-modal sensor data. It is necessary to reduce feature dimensions to facilitate the learning. We propose an optimal feature selection framework which aims to select feature subsets with maximum discrimination capacity. To further address the information loss issue in feature reduction, we develop a novel learning method, termed generalized PDF projection theorem (GPPT), to reconstruct the distribution in high-dimensional raw data space from the low-dimensional feature subspace.

To support the distributed computations throughout the CPS, it needs a novel computing architecture to offer high-performance computing over multiple spatial and temporal scales and to support Internet of Things for machine-to-machine communications. We develop a hierarchical distributed Fog computing architecture for the next generation CPS. A prototype of such architecture for smart pipeline monitoring is implemented to verify its feasibility in real world applications.

Regarding the applications, we examine false data injection detection in smart grid. False data injection is a type of malicious attack which can threaten the security of energy systems. We examine the observability of false data injection and develop statistical models to estimate underlying system states and detect false data injection attacks under different scenarios to enhance the security of power systems (http://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1508&context=oa_diss).

GENETIC PROGRAMMING TECHNIQUES FOR REGULAR EXPRESSION INFERENCE FROM EXAMPLES

Fabiano Tarlao
ftarlao@gmail.com
Department of Engineering and Architecture, University of Trieste, Italy

MACHINE Learning (ML) techniques have proven their effectiveness for obtaining solutions to a given problem automatically, from observations of problem instances and from examples of the desired solution behaviour. In this thesis we describe the work carried out at the Machine Learning Lab at University of Trieste on several real world problems of practical interest.

The main contribution is the design and implementation of a tool, based on Genetic Programming (GP), capable of constructing regular expressions for text extraction automatically, based only on examples of the text to be extracted as well as of the text not to be extracted. Regular expressions are used in a number of different application domains but writing a regular expression for solving a specific task is usually quite difficult, requiring significant technical skills and creativity. The results demonstrate that our tool not only generates text extractors with much higher effectiveness on more difficult tasks than previously proposed algorithms, it is also human-competitive in terms of both accuracy of the regular expressions and time required for their construction. We base this claim on a large-scale experiment involving more than 1700 users on 10 text extraction tasks of realistic complexity. Thanks to these achievements, our proposal has been awarded the Silver Medal at the 13th Annual "Humies" Award, an international competition that establishes the state of the art in genetic and evolutionary computation.

Moreover, in this thesis we consider two variants of the proposed regular expressions generator, tailored to different application domains(i) an automatic Regex Golf game player, i.e., a tool for constructing a regular expression that matches a given list of strings and that does not match another given list of strings; and, (ii) the identification of Genic Interactions in sentences extracted from scientific papers.

This thesis also encompasses contributions beyond the field of Genetic Programming, including: a methodology for predicting the accuracy of text extractors; a novel learning algorithm able to generate a string similarity function tailored to problems of syntax-based entity extraction from unstructured text; a system for continuous reauthentication of web users based on the observed mouse dynamics; a method for the authentication of an unknown document, given a set of verified documents from the same author; a method for user profiling based on a set of his/her tweets; automatic text generators capable of generating fake reviews for a given

scientific paper and fake consumer reviews for a restaurant. The proposed algorithms employ several ML techniques ranging from Grammatical Evolution to Support Vector Machines, from Random Forests to Recurrent Neural Networks (https://drive.google.com/file/d/ 0B67gF86BZtPLdmNyYzZUNF8wTDg/view?usp=sharing).

Data-driven Study on Two Dynamic Evolution Phenomena of Social Networks: Rumor Diffusion in Online Social Media and Bankruptcy Evolution Among Firms

Shihan Wang
ShihanW@trn.dis.titech.ac.jp
Tokyo Institute of Technology, Japan

THE fast growth of computational technologies and unprecedented volume of data have revolutionized the way we understand our society. While the social network structure is commonly used to conceptualize and describe individuals and collectives in the highly connected world, social network analysis becomes an important means of exploring insights behind this social structure. Social networks usually keep evolving slowly over time. This evolution can become very dramatic when facing external influences, which raised new challenges for scholars in understanding the complex social phenomenon.

In the thesis, we concentrate on the dynamic evolution phenomena of social networks caused by external factors from both interpersonal and inter-organizational perspectives: rumor diffusion in online social media (i.e. interpersonal network) and bankruptcy evolution among the firms (i.e. inter-organizational network). Driven by real big data resources, we applied various computational technologies to explore the behavioral patterns in dynamic social networks and provide implications for solving these social problems.

From the individual perspective, we explore the rumor diffusion phenomenon in online social media (i.e. Twitter in particular). With the extremely fast and wide spread of information, online trending rumors cause devastating socioeconomic damage before being effectively identified and corrected. To fix the gap in real-time situation, we propose an early detection mechanism to monitor and identify rumors in the online streaming social media as early as possible. The rumor-related patterns (combining features of users attitude and network structure in information propagation) are first defined, as well as a pattern matching algorithm for tracking the patterns in streaming data. Then, we analyze the snapshots of data stream and alarm matched patterns automatically based on the sliding window mechanism. The experiments in two different real Twitter datasets show that our approach captures early signal patterns of trending rumors and have a good potential to be used in real-time rumor discovery.

From the organizational perspective, we understand the dynamic evolution phenomenon of inter-firm network emerging from bankruptcy. When the bankruptcy transfers as a chain among trade partners (i.e. firms), it causes serious socioeconomic concerns. Beyond previous studies in statistical analysis and propagation modeling, we focus on one underlying humanrelated factor, the social network among senior executives of firms, and investigate its effects on this social phenomenon. Based on empirical analysis of real Japanese firms data in ten years, an agent-based model is particularly proposed to understand the role of this human factor in two perspectives: the number of social partners and the local interaction mechanism among firms (i.e. triangle structure in inter-firm social network). Using real and artificial datasets, the beneficial effects of a number of social partners are well examined and validated in various simulated scenarios from both micro and macro levels. Our results also indicate the influential strategies to keep firms resilience when facing the bankrupt emergency.

Besides the contributions we made in each research field respectively, our study in this thesis enhances the understanding of dynamic independent and interactive behaviors in complex social phenomena, and provides a good perspective to seek solutions in other computational social problems.

Exploring the Knowledge Curation Process of Online Health Communities

Wanli Xing
wanli.xing@ttu.edu
Texas Tech University, USA.

MORE and more people turn to online health communities for social support to satisfy their health-related needs. Previous studies on social support and online health communities in general have focused on the content of social support and the relationship of social support with other entities using traditional social science methods. Little is known about how social support facilitates the knowledge curation process in an online health community. Moreover, the presence of misinformation in online health communities also calls for research into the knowledge curation process in order to reduce the risk of misinformation. This study uses data mining technologies to analyze around one million posts across 23 online health communities along with 900 post information accuracy data. It aims to reveal how information, through social support, flows between the community users working as a whole to dynamically curate knowledge and further interacts with information accuracy.

Text mining methods was used to analyze the 1 million posts data to characterize information flow among the three user positional roles from a quantitative perspective. The results showed that (i) xperiphery users instead of core users dominate the quantitative information flow to request and receive informational support for knowledge curation in online health communities and (iii) the xpeirphery users showed the best potential for generating new information. Granger causality was then employed to analyze the data mining results to characterize information flow between the three user positional roles from content perspective The results demonstrated that (i) it was the xperihpery users who played a central role in directing the content information flow and development; (ii) however, the xperiphery users were still the least active user group in responding to other user positional roles.

Information flow was further quantified from temporal perspective using Directed Acyclic Graphs. K-means clustering and negative binomial regression were further employed to identify three distinct information flow patterns for users to curate knowledge in online health communities and each with distinct characteristics. Further logistic regression models were built to examine the interaction between the identified information flow patterns with information accuracy. The results showed that (i) information patterns and time had a statistically significant influence on information accuracy, and (ii) information accuracy also showed distinct variation trends between information flow patterns. These findings not only have important implications for social support use, delivery and social support research methodologies but also can inform future online health platform design.
(https://www.researchgate.net/publication/317224472_EXPLORING_THE_COLLECTIVE_KNOWLEDGE_CURATION_PROCESS_OF_ONLINE_HEALTH_COMMUNITIES)

Semantic Similarity Analysis and Application in Knowledge Graphs

Ganggao Zhu
gzhu@dit.upm.es
Universidad Politecnica de Madrid, Spain

THE advanced information extraction techniques and increasing availability of linked data have given birth to the notion of large-scale Knowledge Graph (KG). With the increasing popularity of KGs containing millions of concepts and entities, the research of fundamental tools studying semantic features of KGs is critical for the development of KG-based applications, apart from the study of KG population techniques. With such focus, this thesis exploits semantic similarity in KGs taking into consideration of concept taxonomy, concept distribution, entity descriptions, and categories. Semantic similarity captures the closeness of meanings. Through studying the semantic network of concepts and entities with meaningful relations in KGs, we proposed a novel WPath semantic similarity metric and new graph-based Information Content (IC) computation method. With the WPath and graph-based IC, semantic similarity of concepts can be computed directly and only based on the structural and statistical knowledge contained in KG. The word similarity experiments have shown that the improvement of the proposed methods is statistical significant comparing to conventional methods. Moreover, observing that concepts are usually collocated with textual descriptions, we propose a novel embedding approach to train concept and word embedding jointly. The shared vector space of concepts and words has provided convenient similarity computation between concepts and words through vector similarity. Furthermore, the applications of knowledge-based, corpus-based and embedding-based similarity methods are shown and compared to the task of semantic disambiguation and classification, in order to demonstrate the capability and suitability of different similarity methods in the specific application. Finally, semantic entity search is used as an illustrative showcase to demonstrate the higher level of the application consisting of text matching, disambiguation and query expansion. To implement the complete demonstration of entity-centric information querying, we also propose a rule-based approach for constructing and executing SPARQL queries automatically. In summary, the thesis exploits various similarity methods and illustrates their corresponding applications for KGs. The proposed similarity methods and presented similarity based applications would help in facilitating the research and development of applications in KGs (http://oa.upm.es/47031/).